Prévention des effets discriminatoires potentiels de l'utilisation de l'intelligence artificielle dans les services locaux

Note d'orientation politique

Octobre 2020

ePaństwo Foundation

INTERCULTURAL CITIES
CITÉS INTERCULTURELLES

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Sommaire

Rédigé par Krzysztof Izdebski

Unité des Cités interculturelles
Council of Europe©

Conseil de l'Europe, octobre 2020

Krzysztof Izdebski. Membre du Conseil d'administration et Directeur des politiques de ePaństwo Foundation en Pologne, membre de la Consul Democracy Foundation. Krzysztof Izdebski est avocat spécialisé dans l'accès à l'information publique et la réutilisation de l'information du secteur public. Il est l'auteur de publications sur la liberté de l'information, la technologie, l'administration publique, notamment : « Transparency and Open Data Principles: Why They Are Important and How They Increase Public Participation and Tackle Corruption » et de « alGOVrithms. état des lieux » publié récemment. Il est également l'auteur du rapport sur l'utilisation des algorithmes dans les relations entre les pouvoirs publics et les citoyen-n-es en République tchèque, Géorgie, Hongrie, Pologne, Serbie et Slovaquie.

Cette note d'orientation doit servir de document de référence pour le webinaire organisé dans le cadre du programme des « Cités interculturelles » et de la Fondation Epaństwo le 21 septembre 2020.

1. Introduction

1.1 Contexte

Les municipalités fournissent une large gamme de services publics à leurs citoyen-ne-s et, de plus en plus, ces services s'appuient sur des technologies telles que les outils de prise de décision automatisée (ADM) et les solutions d'intelligence artificielle (IA). Le déploiement des outils informatiques dans les services publics a apporté de nouveaux défis et des risques potentiels de partialité, de préjugés envers certaines catégories de citoyen-ne-s et de discrimination. De tels risques ont été, par exemple, détectés dans le système néerlandais SyRI utilisé par les autorités nationales et locales pour détecter les fraudes au logement ou à la sécurité sociale, les compteurs d'eau intelligents dans plusieurs villes d'Europe ou les applications d'IA utilisées pour le recrutement du personnel.

Certaines villes comme New York ont déjà appliqué des mesures pour prévenir ces irrégularités tandis que d'autres commencent juste à réfléchir à celles qui doivent être prises. Les Cités interculturelles développent des politiques et une expertise en matière d'inclusion sociale et d'égalité, de prévention de la discrimination et de sensibilisation aux grands enjeux de société. Il est également utile pour les autorités responsables de comprendre les biais et les risques potentiels de l'IA, et de s'informer sur les moyens d'atténuer ces risques. L'expérience des villes avancées pourrait aider à construire une IA fiable et éthique.

Le programme « Cités interculturelles » a organisé un webinaire consacré aux défis que l'intelligence artificielle et la prise de décision algorithmique représentent pour les autorités locales, en particulier en ce qui concerne l'anti-discrimination, l'inclusion et la lutte contre les discours de haine. Le webinaire a été préparé et dirigé par Krzysztof Izdebski*, Directeur des politiques de ePaństwo Foundation.

Ce rapport reflète le contenu substantiel du webinaire et doit servir de guide succinct sur la prévention des effets discriminatoires potentiels de l'utilisation de l'intelligence artificielle dans les services locaux.

1.2 Glossaire

Intelligence artificielle (IA) : *Toute technologie de l'information qui exécute des tâches pour lesquelles il faut habituellement faire appel à l'intelligence biologique, comme comprendre le langage parlé, apprendre des comportements ou résoudre des problèmes.*
- Directive sur la prise de décision automatisée (Canada)

L'IA n'est qu'un des types d'algorithme susceptibles d'entraîner un risque de discrimination. Comme le mentionne la Charte des algorithmes pour Aotearoa Nouvelle-Zélande, *les risques et bénéfices associés aux algorithmes sont largement* ***décorrélés des types d'algorithmes*** *utilisés. Les algorithmes très simples peuvent être aussi bénéfiques (ou préjudiciables) que les plus complexes, en fonction de leur contenu, de leur objectif et des destinataires visés par les processus métier concernés.*

Il est donc préférable d'utiliser l'expression ***Système décisionnel automatisé*** *qui, selon la* Directive sur la prise de décisions automatisée (Canada) *comprend toute technologie qui soit informe ou remplace le jugement des décideurs humains. Ces systèmes proviennent de domaines tels que les statistiques, la linguistique et les sciences informatiques, et utilisent des techniques telles que les systèmes basés sur*

des règles, la régression, l'analytique prédictive, l'apprentissage automatique, l'apprentissage en profondeur et les réseaux neuronaux.

Plus simplement, selon l'étude « Algorithmics - The Spirit of Computing » (1987) de David Harel, un algorithme est comparable à une recette de cuisine. Si les ingrédients peuvent être comparés aux données d'entrée et le plat terminé au résultat, plusieurs étapes comme le choix des proportions appropriées au moment opportun ou les méthodes de traitement thermique employées ne sont qu'un algorithme. L'expérience pratique permet de démontrer aisément qu'une erreur pendant la préparation d'un plat peut aboutir à un échec, tant sur le plan du goût que sur celui de la présentation.

Selon *C. Orwat in (2020) Risks of Discrimination through the Use of Algorithms:*

La discrimination *est le fait de traiter des personnes de manière défavorable, illégitime et sans justification par rapport à une caractéristique protégée.* Ces caractéristiques peuvent inclure la « race » ou l'origine ethnique, l'ascendance, le pays d'origine, l'origine ; le genre ; la langue ; les opinions ou points de vue politiques, la religion et les convictions ; le handicap ; l'appartenance syndicale, les caractéristiques ou dispositions génétiques et l'état de santé ; les caractéristiques biométriques ; la vie sexuelle, l'identité ou l'orientation sexuelle.

Pour distinguer la discrimination « traditionnelle » de l'IA/ADM discriminatoire, il convient de prendre en considération plusieurs points.

La discrimination fondée sur les choix personnelles consiste à traiter différemment une personne en se fondant sur les préférences ou aversions personnelles ou préconçues des personnes responsables à l'égard d'un groupe donné de personnes ou sur les aversions ou préférences pour certains produits.

La discrimination statistique consiste à traiter différemment des personnes, de manière illégitime et injustifiée, en se fondant sur des informations de substitution.

Il est cependant primordial de comprendre que les algorithmes étant créés par des êtres humains avec tous les biais que cela suppose, la discrimination statistique peut être la conséquence d'une discrimination fondée sur les préférences personnelles. Ces deux phénomènes sont donc très rarement distincts l'un de l'autre.

2 Conséquences et exemple de discrimination algorithmique

Un certain nombre d'exemples montrent les effets discriminatoires des outils d'IA/ADM. Citons à ce propos le problème posé récemment par l'algorithme de notation du A-Level, l'équivalent anglais du baccalauréat au Royaume-Uni. Les chiffres ont montré que 39,1 % des notes attribuées par les 700 000 enseignant-e-s ont été inférieures d'au moins un point, cette différence étant particulièrement accentuée parmi les élèves des groupes socio-économiques les moins favorisés.

A-level results: almost 40% of teacher assessments in England downgraded

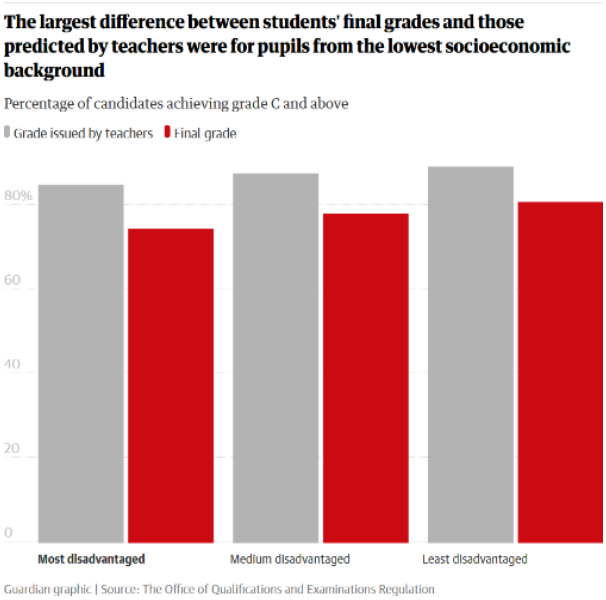
Ofqual figures show 39.1% of 700,000 teacher assessments were lowered by at least one grade

A-level results - live updates

2:09

'I put my heart and soul into them': A-level students on downgraded results - video report

Teachers in England had nearly 40% of their A-level assessments downgraded by the exam regulator's algorithm, according to official figures published on Thursday morning as sixth-formers around the UK received their results.



The largest difference between students' final grades and those predicted by teachers were for pupils from the lowest socioeconomic background

Percentage of candidates achieving grade C and above

Grade issued by teachers | Final grade

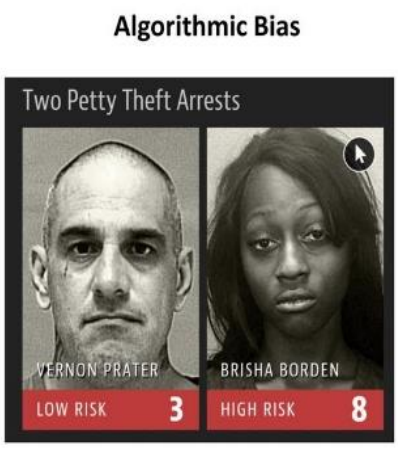
80%
60
40
20
0

Most disadvantaged | Medium disadvantaged | Least disadvantaged

Guardian graphic | Source: The Office of Qualifications and Examinations Regulation

Selon certains juges, ces problèmes s'expliquent le plus souvent par le fait que *les décisions automatisées n'incluent pas d'évaluation détaillée des circonstances spécifiques du cas en question. Un fonctionnaire pourrait au contraire mieux motiver sa décision, et les arguments sur lesquels elle repose, et ce faisant, mieux délimiter le champ des contestations éventuelles. Le contexte est essentiel pour éviter de prendre involontairement des décisions biaisées.*

Eric Holter, ancien procureur général des États-Unis adopte une démarche similaire lorsqu'il affirme à propos de la détermination des peines fondée sur des algorithmes que, *même si ces mesures ont été conçues dans les meilleures intentions, je suis préoccupé par le fait qu'elles sapent involontairement nos efforts afin de garantir l'égalité et l'individualisation de la justice.* Cette déclaration a été effectuée après le scandale provoqué par COMPAS, un outil d'IA/ADM utilisé aux États-Unis pour prédire la probabilité de commettre un crime à l'avenir. Les affaires Brisha Borden et Vernon Prater ont révélé que *les technologies décisionnelles fondées sur des données utilisées dans le système judiciaire pour éclairer des décisions relatives à la libération sous caution, la mise en liberté conditionnelle et aux peines d'emprisonnement sont entachées de biais à l'encontre des groupes historiquement marginalisés*.



En outre, la Commission européenne considère que *certains algorithmes d'IA peuvent, lorsqu'ils sont utilisés pour* ***prédire la récidive d'actes délictueux,*** *présenter des biais de nature sexiste et raciale et fournissent des prédictions de la probabilité de récidive différentes selon qu'il s'agit de femmes ou d'hommes, ou de ressortissants nationaux ou d'étrangers.* Un autre exemple fait référence à *certains programmes d'IA pour* ***la reconnaissance faciale*** *qui sont entachés de biais de nature sexiste ou raciale, lesquels se traduisent par un faible taux d'erreur dans la détermination du sexe des hommes à peau claire, mais un taux d'erreur élevé dans la détermination du sexe des femmes à peau foncée.*

Les participantes et participants ont ainsi été amenés à approfondir la question de la discrimination statistique en se basant sur l'exemple suivant.

> En Belgique, un prestataire de service d'énergie refuse d'approvisionner en électricité les personnes qui vivent dans un quartier précis. Il estime qu'un grand nombre d'habitant-e-s de ce quartier sont des mauvais payeurs. Même les client-e-s potentiels solvables sont exclus de la fourniture d'électricité sans tenir compte de leur niveau de solvabilité individuelle.

Dans ce cas, la **variable de substitution** est le « lieu de résidence ».

3 Comment fonctionne la discrimination dans le domaine de l'IA

En s'appuyant sur le rapport de F. Z. Borgesius (2018) « *Discrimination, artificial intelligence, and algorithmic decision-making* », Krzysztof Izdebski explique comment l'IA/ADM peut conduire à différentes formes de discrimination :

> (i) Comment sont définis la « variable cible » et les « étiquettes de classe » ; (ii) l'étiquetage des données d'apprentissage ; (iii) la collecte des données d'apprentissage (iv) la sélection des caractéristiques ; et (v) les données indirectes ainsi que (vi), les systèmes d'AI peuvent être utilisés, à dessein, à des fins discriminatoires.

Variable cible et étiquettes de classe *« l'algorithme d'apprentissage automatique tire d'exemples pertinents (fraudes précédemment identifiées, courriers indésirables, défauts de paiement, mauvaise santé) les attributs ou les actions (données indirectes) qui peuvent servir à détecter la présence ou l'absence de la qualité ou du résultat recherchés (la variable cible) ». Ce résultat recherché est appelé « variable cible ».* Les étiquettes de classe sont liées aux variables cibles. *Prenons le cas d'une entreprise qui confie à un système d'IA le soin de classer des réponses à une offre d'emploi pour en extraire de « bons employés ». Comment va-t-on définir le bon employé ? En d'autres termes, quelles devraient être les « étiquettes de classe » ? Le bon employé est-il celui qui réalise les meilleures ventes ou celui qui n'arrive jamais en retard au travail ?* Selon Borgesius, *la discrimination peut s'introduire dans un système d'IA en raison de la façon dont une organisation définit les variables cibles et les étiquettes de classe.*

Étiquetage des données d'apprentissage *Un système d'IA peut faire son apprentissage sur des données biaisées [ou] des problèmes peuvent surgir si le système apprend à partir d'un échantillon biaisé.* Borgesius donne des exemples de systèmes créés pour trier des demandes d'inscription à l'université. *Les don-*

nées d'apprentissage du programme informatique étaient les dossiers d'admission des années précédentes qui comportaient des biais de genre et d'appartenance ethnique, ce qui a abouti à un nombre moins élevé d'inscriptions de femmes et de personnes issues de la migration.

Collecte des données d'apprentissage *La procédure d'échantillonnage peut aussi être biaisée. Par exemple, dans la collecte de données sur la criminalité, il se pourrait que la police ait interpellé dans le passé plus de personnes issues de la migration. Lum et Isaac observent que si la police se concentre sur certains groupes ethniques et certains quartiers, il est probable que ces catégories seront surreprésentées dans ses fichiers.*

Sélection des caractéristiques *Supposons qu'une organisation veuille sélectionner par prédiction automatisée les candidats qui seront de bons employés. Il est impossible, ou du moins trop coûteux, pour un système d'IA d'évaluer la totalité de chaque dossier de candidature. L'organisation peut alors retenir, par exemple, uniquement certaines caractéristiques applicables à chaque dossier. Le choix de certaines caractéristiques peut introduire un biais contre certains groupes.*

Données indirectes : *Certaines données incluses dans le jeu d'apprentissage peuvent présenter des corrélations avec des caractéristiques protégées.* [...] *Les données d'apprentissage ne contiennent pas d'informations concernant des caractéristiques protégées, comme la couleur de la peau. Le système d'IA apprend que les personnes qui ont un certain code postal ont tendance à ne pas rembourser, et il utilise cette corrélation pour prédire le non-remboursement du crédit. Un critère à première vue neutre (le code postal) sert donc à prédire le défaut de paiement. Mais supposons maintenant qu'il y ait une corrélation entre ce code postal et l'origine raciale.*

Si la banque prend ses décisions sur la base de cette prédiction et refuse d'accorder des crédits aux habitants de ce quartier, cela fait du tort aux membres d'un certain groupe sur le critère de l'origine raciale. L'organisation peut aussi utiliser intentionnellement les données indirectes pour pratiquer la discrimination sur la base de l'origine raciale.

4 Comment prévenir la discrimination dans les outils d'AI/ADM

Certaines méthodes peuvent contribuer à lutter contre le risque de discrimination ou le réduire au minimum lors de l'utilisation d'outils d'IA/ADM.

On dispose ainsi de **solutions centrées sur l'humain** intégrées aux procédures de marchés publics et des **évaluations d'impact algorithmique**.

Les Principes directeurs applicables aux marchés publics du Forum économique mondial contiennent les **10 principes suivants pour prévenir les biais ou préjugés liés à l'IA/ADM.**

Une IA/ADM « digne de confiance » selon le Groupe d'experts indépendants de haut niveau de la Commission européenne respecte les principes suivants : action humaine et contrôle humain ; robustesse technique et sécurité ; respect de la vie privée et gouvernance des données ; transparence, diversité, non-discrimination et équité ; bien-être sociétal et environnemental, et responsabilité. Ces principes doivent être respectés lors de la planification des marchés publics.

1. Utilisez des procédures de marchés publics qui ne sont pas axées sur la recherche d'une solution spécifique, mais plutôt sur la description des problèmes et opportunités et favorisez une approche itérative.
2. Définissez les bénéfices de l'utilisation de l'IA pour le public tout en évaluant les risques.
3. Alignez vos passations de marchés sur les stratégies des pouvoirs publics en la matière et contribuez à leur amélioration.
4. Intégrez la législation et les codes de bonnes pratiques potentiellement applicables dans vos demandes de propositions.
5. Examinez la faisabilité technique et administrative de l'accès aux données utiles.
6. Soulignez les limites techniques et éthiques de l'utilisation qui sera faite des données afin d'éviter des problèmes tels que les biais de données historiques.
7. Travaillez avec une équipe multidisciplinaire et diversifiée.
8. Privilégiez, à toutes les étapes du mécanisme de passation, la responsabilité des algorithmes et les normes de transparence.
9. Mettez en œuvre un processus d'échange constant entre le prestataire d'IA et l'entité qui l'acquiert, à des fins de transfert de connaissances et d'évaluation des risques à long terme.
10. Placez les prestataires de solutions d'IA dans des conditions égales et équitables.

Ainsi, pour garantir la transparence du mécanisme, l'une des conditions décrites dans l'avis de marché pourrait comprendre une solution en libre accès (« open source »), permettant à des expert-e-s externes d'analyser le code logiciel pour signaler des risques potentiels de corruption.

Un **exemple pratique** d'application d'une Évaluation de l'impact algorithmique est présenté dans la matrice des risques de la Charte des algorithmes pour Aotearoa Nouvelle-Zélande. Les **Principaux éléments de l'Évaluation de l'impact algorithmique (EIA) d'un organisme public,** décrits dans l'étude « Algorithmic Impact Assessments : A Practical Framework For Public Agency Accountability » de l'Institut AI Now, sont présentés ci-après.

Ce tableau doit être utilisé avant de répondre au questionnaire pratique sur l'EIA qui permet de cerner plus précisément les risques et est utilisé pour décrire l'impact concret en termes de discrimination. Un exemple du Canada a été présenté aux participant-e-s. Le questionnaire EIA contient des questions comme :

▪ La recommandation ou la décision formulée par le système contient-elle des éléments d'appréciation ? ▪ - Décrivez ce qui est discrétionnaire dans la décision. ▪ - Le système est-il utilisé par une partie de l'organisation qui n'est pas celle qui l'a développé ? ▪ - Les impacts résultant de la décision sont-ils réversibles ?	▪ La recommandation ou la décision formulée par le système comprend-elle des éléments d'appréciation ? ▪ - Le système est-il utilisé par une partie de l'organisation qui n'est pas celle qui l'a développé ? ▪ Les impacts résultant de la décision sont-ils réversibles ? ▪ Quelle sera la durée des impacts de la décision ?	▪ Le Système décisionnel automatisé utilise-t-il des informations comme données d'entrée ? ▪ Quelle est la classification de sécurité la plus élevée pour les données d'entrée utilisées par le système ? (Faire un choix) ▪ Qui contrôle les données ? ▪ Qui a recueilli les données utilisées pour l'apprentissage du système ? ▪ Qui a recueilli les données utilisées par le système ?

1. Les organismes publics devraient procéder à une auto-évaluation des systèmes décisionnels automatisés existants et prévus, en évaluant leurs impacts potentiels sur l'équité, la justice, la partialité ou d'autres préoccupations au sein des communautés concernées.

2. Les organismes publics devraient mettre en place des processus d'examen efficaces confiés à des chercheur-e-s externes afin d'observer, mesurer ou suivre les impacts dans le temps ;

3. Les organismes publics devraient informer le public de leur définition d'un « système décisionnel automatisé », des systèmes existants et prévus ainsi que des éventuels processus d'auto-évaluation et d'examen par des chercheur-e-s avant l'acquisition du système ;

4. Les organismes publics devraient solliciter les avis du public pour apporter des éclaircissements et répondre aux questions en suspens ; et

5. Les pouvoirs publics devraient prévoir des dispositifs de recours améliorés afin que les communautés ou individus concernés puissent contester des évaluations inappropriées ou l'utilisation d'un système inéquitable, biaisé ou autrement préjudiciable que les organismes publics ne sont pas parvenus à modérer ou corriger.

Risk matrix

Likelihood

Probable Likely to occur often during standard operations			
Occasional Likely to occur some time during standard operations			
Improbable Unlikely but possible to occur during standard operations			
Impact	***Low*** The impact of these decisions is isolated and/or their severity is not serious.	***Moderate*** The impact of these decisions reaches a moderate amount of people and/or their severity is moderate.	***High*** The impact of these decisions is widespread and/or their severity is serious.

Risk rating

Low The Algorithm Charter could be applied.	***Moderate*** The Algorithm Charter should be applied.	***High*** The Algorithm Charter must be applied.

5 Résumé

Pour prévenir les risques potentiels de discrimination, les municipalités qui veulent se préparer à une utilisation élargie des solutions d'IA/ADM doivent :

- adopter des politiques sur la mise en œuvre des algorithmes, qui décrivent le processus et les personnes responsables (de préférence une équipe pluridisciplinaire et diversifiée) ;
- mettre en place des Évaluations de l'impact algorithmique ;
- introduire des clauses de transparence dans les contrats avec les entreprises qui fournissent des logiciels et un accès ouvert au code source, dans le grand public ou tout au moins parmi les experts externes ;
- publier des lignes directrices expliquant le fonctionnement des algorithmes aux personnes directement concernées ;
- développer davantage le système d'analyse des solutions d'IA/ADM, en faisant, là encore, participer une équipe pluridisciplinaire et diversifiée ;
- associer citoyens et experts à la planification de la passation de marchés et à la mise en œuvre de l'IA/ADM, et permettre ainsi d'identifier des risques potentiels de discrimination ;
- intégrer l'utilisation des solutions d'IA/ADM dans les programmes de renforcement des connaissances et des compétences des agents publics et des autres employés municipaux directement ou indirectement concernés par leur utilisation.